技術報告　「令和 2 年度 科学研究費助成事業(奨励研究)」
深層強化学習による適応的制御
～並列 RN の導入による多様な時間スケールへの適応～

松木俊貴

大分大学理工学部技術部

目的

深層学習の手法によって学習したニューラルネットワーク (NN) による認識システムが，既存のシステムを凌駕する性能を発揮し注目を集めている．また，明示的な教示なしに，探索と報酬を通じて学習を行う強化学習と深層学習とを組み合わせた深層強化学習の研究が盛んに行われている．生活のあらゆる場面で活躍し，人と協働できるロボットの実現のため，新しい状況でも適応可能で，目標達成のために非線形に動きを変えていけるような制御が必要だと考える．そのような制御が，深層強化学習により実現できないかと考え，これまで，時系列データ処理を高速に学習することが可能なリザバーネットワーク (RN) を導入した深層強化学習について研究を行ってきた．その中で，即応的な反応と長期的な運動を同時に学習することの難しさが一つの課題として浮かび上がった．そこで異なる時間スケールに対応する RN を並列に構成したネットワーク構造を導入し，強化学習によって複数時間スケールの反応が必要なタスクを学習することで，その手法の有効性を調査した．

研究方法

図 1 に学習の全体構成を示す．本研究では，強化学習用ライブラリ gym に含まれる Cart-Pole タスクをもとにして，複数時間スケールのタスクを開発した．学習エージェントは通常の振り子とその3倍の長さを持つ振り子が取り付けられたカートを動かし，できるだけ長い間両方の振り子を立たせることを目的とする．長い振り子はゆっくりとした時間スケールで，短い振り子は小刻みに動く早い時間スケールでの処理が求められる．このタスクでは，どちらかの振り子が一定以上傾くかカートの位置が一定の範囲を超える，もしくは上限の200ステップが経過するまでを1試行とし，規定の範囲で振り子を立たせている間，毎ステップ1の報酬が環境からエージェントへ与えられる．エージェントは環境からカートの位置と速度，二つの振り子の角度と角速度の情報を入力として受け取り，カートを左右どちらに押すかの行動選択をする．

図 1 に示すように，エージェントネットワークは二つの RN と出力を生成する NN から構成される．RN は環境からの入力を受け取りその情報を内的なダイナミクスとして時空間的に展開した出力を生成する．NN は，RN の出力を入力として受け取り行動選択を決定する．RN 内部のニューロンは差分近似微分方程式に従って動作しており，そのダイナミクスの変化速度は時定数τと呼ばれるパラメータによって変わる．ここでは簡単のためにシミュレーションの時間間隔を 1 としているため，τは1が最小値となり，その値が大きくなるにつれて RN のダイナミクスはゆっくりと変化するようになる．すなわち RN の状態は，τ が1だと入力に対して即応的に変化するようになり，τ が大きくなるほど入力に対して長期的な反応を示すようになる．ここで，一方の RN は，τの値が1に設定されており，もう一つの RN は任意の時定数τ_sに設定される．このように異なる時定数を持つ RN を並列して構成したネットワークをこれ以降混合モデルと呼ぶ. RN 内部のニューロン数はそれぞれ100個としそれぞれ結合確率10%で相互結合している．相互結合重み値は，重み値行列のスペクトル半径が 0.97 になるように決定した．行動選択を行う NN は200個のニューロンで構成される中間層を1層持つネットワークを用いた．

ネットワークの学習は Deep Q-Network と呼ばれる深層強化学習の手法に基づいて行なった．探索は ε-greedy 法によって行い，ε の初期値は1で1800試行の時点で0.05になるように一定の比率で等比的に減衰させ，1800試行以降は0.05で固定した．また，30000ステップ分の経験をリプレイメモリに蓄え，バッチサイズ500でバッチ学習を行う経験リプレイ学習を行った．NN は SGD により学習を行い，学習係数は初期値0.01から最終試行で0.001になるように一定の値を引いていった．

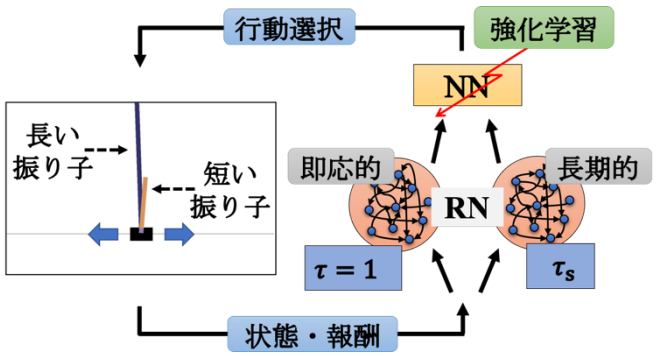

図 1 学習の全体構成

結果

図 2 に 5000 試行学習を行い，その後探索と学習を停止して 100 試行の間テストを行った際の学習曲線を示す．縦軸は各試行で獲得した合計報酬の値であり，横軸は試行回数である．このデータは 20 個の乱数系列を使って学習したもののうち，学習に成功したものから10系列分のデータを取り出し平均して算出した．また，200ステップで移動平均をとっている．各グラフは時定数τ_sの値を$1, 3, 5, 7$とした場合の結果を示している．図から，τ_sをどの値に設定した場合も上限となる200ステップに近い期間振り子を立たせ続けることの学習に成功していることがわかる．一方図 3 に，τ_sを1から10まで1刻みで変化させ，それぞれの条件で乱数系列20パターンの学習を行なった時の学習成功率を示す．ここで，テストフェーズの100試行の間振り上げられたステップ数の平均が180以上であったケースを成功としてカウントしている．図からわかるように，$\tau_s = 1$の時，すなわち両方のリザバが即応的なダイナミクスをもっていた場合に学習成功確率は55%であった．そしてその一方で，τ_sを1より大きくしてゆっくりと変化する RN を並列させた混合モデルの方が常に学習成功確率が向上していることがわかる．

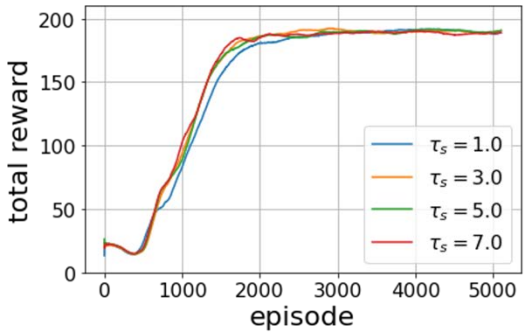

図 2 学習曲線

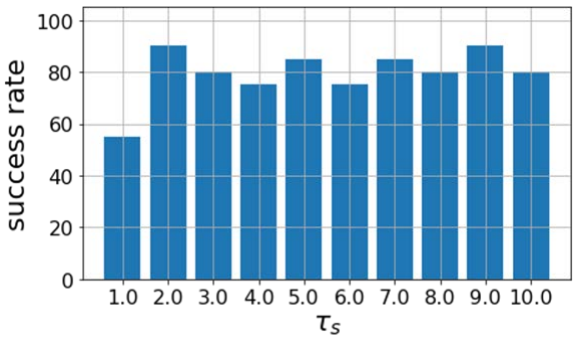

図 3 学習成功確率

しかし，これらの結果だけでは，単に1より大きい時定数を持つ RN の方がここで使用したタスクの学習に有利であったという可能性を否定できない．そこで，異なる時定数の RN を並列させた混合モデルの有効性を確認するため，両方の RN の時定数を任意の値τ_sで同様の値にした非混合モデルを使い，τ_sを1から変化させて行った時に学習性能がどのように変わるかを確かめた．図 4 はこの時の学習曲線を示している．図から$\tau_s = 1$と$\tau_s = 3$の場合は学習曲線に大きな変化はないものの，それ以上に大きくした場合に学習性能が低下していることがわかる．また図 5 に，この場合のτ_sごとの学習成功確率を示した．ここで，オレンジのバーが非混合モデルを使った場合の結果であり，比較のために図 3 に示したものと同じデータを青いバーで示した．この図から，$\tau_s = 4$までは，両者の学習成功確率は同値かもしくは混合モデルの方がわずかに高くなっている．また，それ以上にτ_sを大きくしていくと学習成功確率が低下していくことが分かる．

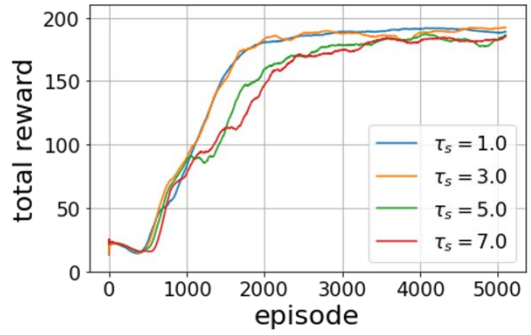

図 4 学習曲線（非混合モデル）

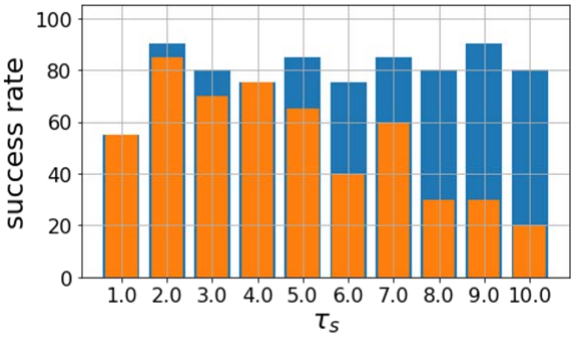

図 5 学習成功確率（非混合モデル）

まとめと今後の課題

以上の結果から，時定数の異なる RN を並列に構成することで，複数の時間スケールを持つタスクの学習性能が向上する可能性を示唆することができた．一方，このような結果がどれだけのタスクに対し汎用的に成り立つかはまだ明らかではない．今後，さらに多くの学習対象のタスクを用いて学習性能を調べ，提案手法の有効性を確かめていく必要がある．

謝辞

本研究は JSPS 科研費（奨励研究）JP20H01158 の助成を受けた．